# Latent Semantic Analysis (LSA)

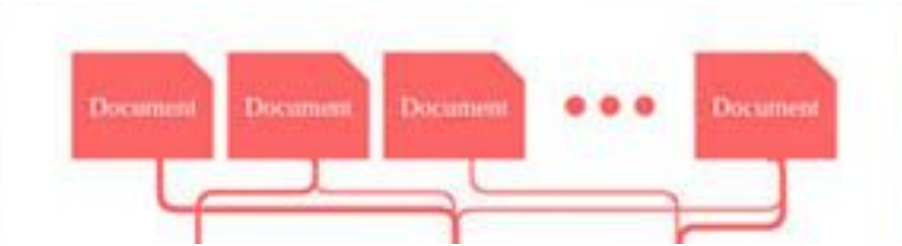
KE_Team 2

# Nội dung

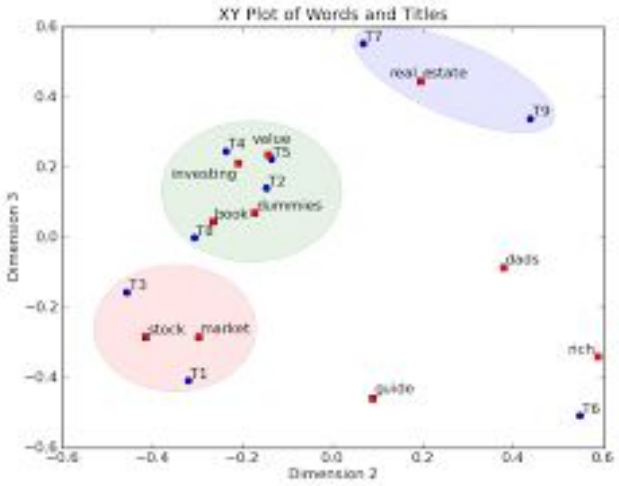
1. Giới thiệu tổng quan
2. Latent Semantic Analysis
3. Triển khai
4. Nhận xét, đánh giá

# Giới thiệu tổng quan

Thu thập thông tin từ lượng lớn văn bản ?




Biểu diễn tri thức như thế nào ?



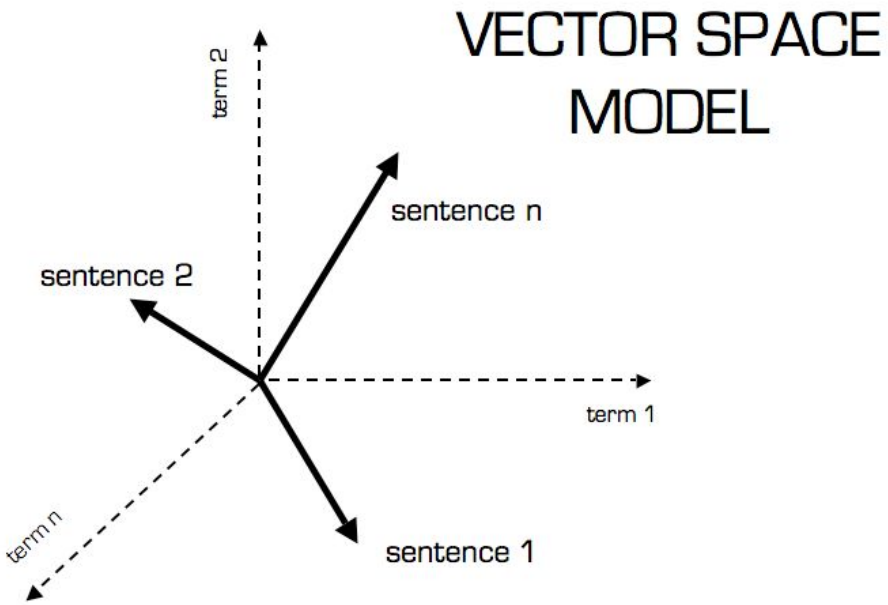

Sự liên quan/ tương đồng giữa dữ liệu ?

# Phân tích ngữ nghĩa tiềm ẩn (LSA)

+ Nền tảng toán học
+ Biểu diễn tri thức
+ Cách sử dụng
+ Demo thuật tonas

# Biểu diễn không gian vector

- Biểu diễn đầu vào dưới dạng các vector
- Các chiều là các features
- Tính toán dựa trên biểu diễn vector



# Term-Document Matrix

Example Documents (Corpus)

- d1 : Romeo and Juliet.
- d2 : Juliet: O happy dagger!
- d3 : Romeo died by dagger.
- d4 : “Live free or die”, that’s the New-Hampshire’s motto.
- d5 : Did you know, New-Hampshire is in New-England.

| | $d_1$ | $d_2$ | $d_3$ | $d_4$ | $d_5$ |
|---|---|---|---|---|---|
| *romeo* | 1 | 0 | 1 | 0 | 0 |
| *juliet* | 1 | 1 | 0 | 0 | 0 |
| *happy* | 0 | 1 | 0 | 0 | 0 |
| *dagger* | 0 | 1 | 1 | 0 | 0 |
| *live* | 0 | 0 | 0 | 1 | 0 |
| *die* | 0 | 0 | 1 | 1 | 0 |
| *free* | 0 | 0 | 0 | 1 | 0 |
| *new-hampshire* | 0 | 0 | 0 | 1 | 1 |

Example of Document Term Matrix

# Singular Value Decomposition

Document-Term Matrix rank r (8x5)

| | | | | |
|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 0 |
| 1 | 1 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 | 0 |
| 0 | 1 | 1 | 0 | 0 |
| 0 | 0 | 0 | 1 | 0 |
| 0 | 0 | 1 | 1 | 0 |
| 0 | 0 | 0 | 1 | 0 |
| 0 | 0 | 0 | 1 | 1 |

$$A = U\Sigma V^T$$

| | | | | |
|---|---|---|---|---|
| -0.396153 | 0.280057 | -0.571171 | 0.449685 | -0.101839 |
| -0.314268 | 0.449532 | 0.410591 | 0.513018 | 0.203906 |
| -0.17824 | 0.268992 | 0.497321 | -0.256998 | 0.0430523 |
| -0.438364 | 0.368508 | 0.0128792 | -0.577329 | -0.21964 |
| -0.263881 | -0.345921 | 0.145789 | 0.0474849 | 0.417484 |
| -0.524005 | -0.246405 | -0.338652 | -0.272846 | 0.154791 |
| -0.263881 | -0.345921 | 0.145789 | 0.0474849 | 0.417484 |
| -0.326373 | -0.459669 | 0.317003 | 0.237244 | -0.724851 |

Left Singular Vectors (8x5)

| | | | | |
|---|---|---|---|---|
| 2.2853 | · | · | · | · |
| · | 2.01026 | · | · | · |
| · | · | 1.3607 | · | · |
| · | · | · | 1.11814 | · |
| · | · | · | · | 0.796577 |

Singular Values (5x5)

| | | | | |
|---|---|---|---|---|
| -0.310866 | -0.40733 | -0.594461 | -0.603046 | -0.142814 |
| 0.362933 | 0.540742 | 0.200054 | -0.695391 | -0.228662 |
| -0.118013 | 0.676704 | -0.659179 | 0.198375 | 0.232971 |
| 0.860986 | -0.28736 | -0.358175 | 0.0530948 | 0.212177 |
| 0.128132 | 0.0342945 | -0.209255 | 0.332558 | -0.909958 |

Right Singular Value (5x5)

# Xấp xỉ low-rank

## Low-rank Approximations

Xấp xỉ của ma trận A

$$\tilde{A} = \underset{X:rank(X)=k}{Min} \|A - X\|_F$$

**Frobenius norm**

$$\|A\| = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2}$$

Nghiệm tối ưu Truncated SVD

$$A_k = U_k \Sigma_k V_k^T$$

# Truncated SVD

New Matrix - rank k < r

| | | | | |
|---|---|---|---|---|
| 0.485762 | 0.673199 | 0.65081 | 0.154457 | 0.00056003 |
| 0.551237 | 0.781199 | 0.607724 | -0.195303 | -0.104067 |
| 0.322878 | 0.45832 | 0.35032 | -0.130389 | -0.0654744 |
| 0.580283 | 0.808641 | 0.743726 | 0.0889825 | -0.0263216 |
| -0.0649144 | -0.130389 | 0.219371 | 0.847233 | 0.245133 |
| 0.19249 | 0.219931 | 0.612777 | 1.0666 | 0.284286 |
| -0.0649144 | -0.130389 | 0.219371 | 0.847233 | 0.245133 |
| -0.103507 | -0.195863 | 0.258524 | 1.09237 | 0.317815 |

$$A_k = U_k \Sigma_k V_k^T$$

Demo: k = 2

| | | | | |
|---|---|---|---|---|
| -0.396153 | 0.280057 | -0.571171 | 0.449685 | -0.101839 |
| -0.314268 | 0.449532 | 0.410591 | 0.513018 | 0.203906 |
| -0.17824 | 0.268992 | 0.497321 | -0.256998 | 0.0430523 |
| -0.438364 | 0.368508 | 0.0128792 | -0.577329 | -0.21964 |
| -0.263881 | -0.345921 | 0.145789 | 0.0474849 | 0.417484 |
| -0.524005 | -0.246405 | -0.338652 | -0.272846 | 0.154791 |
| -0.263881 | -0.345921 | 0.145789 | 0.0474849 | 0.417484 |
| -0.326373 | -0.459669 | 0.317003 | 0.237244 | -0.724851 |

8x2

2x2

| | | | | |
|---|---|---|---|---|
| 2.2853 | . | . | . | . |
| . | 2.01026 | . | . | . |
| . | . | 1.3607 | . | . |
| . | . | . | 1.11814 | . |
| . | . | . | . | 0.796577 |

2x5

| | | | | |
|---|---|---|---|---|
| -0.310866 | -0.40733 | -0.594461 | -0.603046 | -0.142814 |
| 0.362933 | 0.540742 | 0.200054 | -0.695391 | -0.228662 |
| -0.118013 | 0.676704 | -0.659179 | 0.198375 | 0.232971 |
| 0.860986 | -0.28736 | -0.358175 | 0.0530948 | 0.212177 |
| 0.128132 | 0.0342945 | -0.209255 | 0.332558 | -0.909958 |

# Không gian LSA

Term Vectors Matrix

$$U_2\Sigma_2$$

| | | |
|---|---|---|
| *romeo* | -0.905327 | -0.562988 |
| *juliet* | -0.718196 | -0.903676 |
| *happy* | -0.40733 | -0.540742 |
| *dagger* | -1.00179 | -0.740797 |
| *live* | -0.603046 | 0.695391 |
| *die* | -1.19751 | 0.495337 |
| *free* | -0.603046 | 0.695391 |
| *new-hampshire* | -0.74586 | 0.924053 |

Document Vectors Matrix

$$\Sigma_2 S_2^T$$

| | | |
|---|---|---|
| *d1* | -0.710421 | 0.72959 |
| *d2* | -0.930871 | 1.08703 |
| *d3* | -1.35852 | 0.402161 |
| *d4* | -1.37814 | -1.39792 |
| *d5* | -0.326373 | -0.459669 |

# Không gian LSA

- Query q = [q1, q2, ...]

$$q = \frac{\sum_{i=1}^{N} q_i}{N}$$

Centroid

Ví dụ: Query q = [die, dagger]

die = [-1.197  -0.494]
dagger = [-1.001  0.742]

q = ([-1.197  -0.494] + [-1.001  0.742]) / 2
  = [-1.099  0.124]

- Độ tương đồng giữa q và document d

$$s = \frac{q.d}{|q||d|}$$

# Updating/ Downdating

Out of Vocabulary (OOV)? New Documents?

Folding-in k-dimension representation

$$t_n = t.V_k.\Sigma_k^-1$$
$$d_n = d^T.U_k.\Sigma_k^-1$$

=> LSA có khả năng biểu diễn những từ/ văn bản mới dù chưa từng gặp

# Triển khai

Áp dụng mô hình vào dữ liệu thực tế

1. Tiền xử lý dữ liệu
2. Lựa chọn mô hình
3. Biểu diễn trực quan

# Chuẩn bị dữ liệu

Nguồn dữ liệu:
A Large-scale Vietnamese News Text Classification Corpus (https://github.com/duyvuleo/VNTC)

10Topics/Ver1.1/Train_Full

- Chính trị - xã hội
- Đời sống
- Khoa học
- Kinh doanh
- Pháp luật
- Sức khỏe
- Thế giới
- Thể thao
- Văn hoá
- Vi tính

=> 33 759 files, 143.5 MiB

'Hà Nội sắp ngừng đăng ký xe máy tại 3 quận \n Theo Phó chủ tịch UBND thành phố Đỗ Hoàng Ân, đầu năm 2005 có thể tiếp tục hạn chế đăng ký xe máy mới tại 3 quận Cầu Giấy, Thanh Xuân, Tây Hồ theo lộ trình. Hiện các ban ngành thành phố đang nghiên cứu để đưa ra thời điểm chính thức. \n- Thưa ông, tại sao phải ngừng đăng ký xe máy thêm 3 quận nội thành? \n- Theo Nghị quyết HĐND, thành phố sẽ ngừng đăng ký xe máy mới tại các quận theo lộ trình nhằm hạn chế phương tiện cá nhân. Hiện nay, quản lý của nhà nước chưa rõ, phương tiện do người dân mua nhiều gây rối loạn thị trường và ùn tắc giao thông. \n- Hạn chế xe mới ở nội thành thì có thể xe ngoại tỉnh tràn vào thành phố, có ý kiến nên hạn chế cả xe ngoại tỉnh, ông nghĩ sao?\n- Không thể hạn chế xe ngoại tỉnh vì người dân sống trong và ngoài thành phố vẫn sử dụng phương tiện theo nhu cầu cuộc sống. \n Song có lẽ trong thành phố nghĩ đến hạn chế bớt phương tiện vào khu vực trung tâm, tạo ra những tuyến chuyên dành cho xe buýt để hạn chế giao thông cá nhân. Như ở một một số nước, phương tiện cá nhân nào sử dụng thời gian lâu ở trong nội thành sẽ có khoản phí nhất định, sẽ hạn chế xe cá nhân và thúc đẩy phương tiện công cộng. ...

Văn bản mẫu

# Xử lý dữ liệu

1. Tách văn bản thành các tokens để xử lý
2. Loại bỏ các từ không hợp lệ
3. Loại bỏ các stopwords
4. Tạo 1 bộ từ điển idx2term
5. Tạo term-document matrix từ bộ dữ liệu

```
'ubnd',
'thanh xuân'
'ban ngành',
'nghiên cứu'
'nghị quyết'
'mua',
'giao thông'
'ngoại tỉnh'
```

Tách tokens

```
61732 words
0 : ban ngành
1 : cao tầng
2 : chuyên
3 : duy trì
4 : gia tăng
5 : giao
6 : giao thông
7 : italy
8 : khoản
9 : khu vực
10 : khuyến
11 : kinh nghiệm
12 : mua
13 : nghiên cứu
14 : nghị quyết
15 : ngoại tỉnh
16 : nguy hiểm
17 : nguyên nhân
18 : nhu cầu
19 : quy hoạch
```

Từ điển từ

```
['a lô', 'a ha', 'ai', 'ai ai', 'ai nấy']
```

Stopwords trong tiếng Việt

```
[(0, 1), (1, 1), (2, 1), (3, 1), (4, 1), (5, 1), (6, 8), (7, 1), (8, 1), (9, 1), (10, 1), (11, 1), (12, 1), (13, 1),
(14, 1), (15, 3), (16, 1), (17, 1), (18, 4), (19, 2), (20, 1), (21, 7), (22, 5), (23, 1), (24, 1), (25, 1), (26, 1),
(27, 1), (28, 1), (29, 1), (30, 1)]
```

Biểu diễn doc[0] ở term-doc matrix

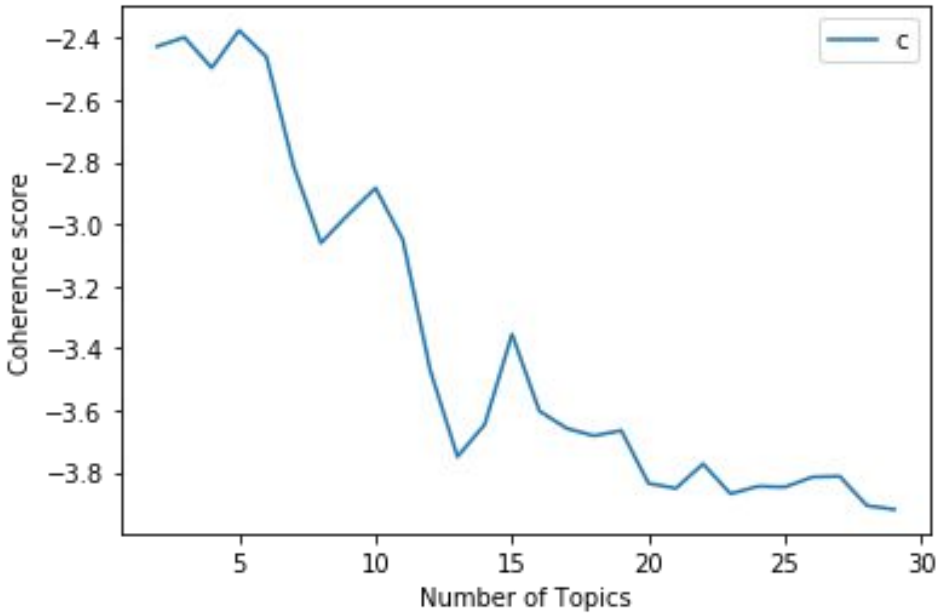
# Xây dựng mô hình

- Sử dụng gensim.LsiModel
- Đầu vào :
  - idx2term dictionary
  - term-document matrix

Lựa chọn số lượng topic như thế nào ?

- gensim.CoherenceModel
- coherence score = 'u_mass'
- Thử từ topic = 2 -> 30

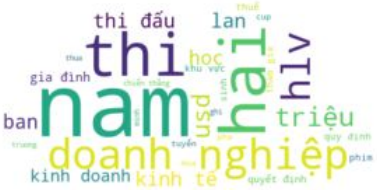
=> Best : 5 topic



# Biểu diễn trực quan

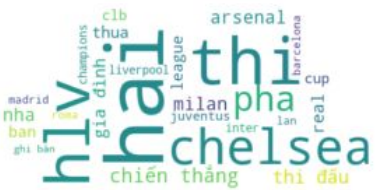
Visualization

Topic 1




Topic 2

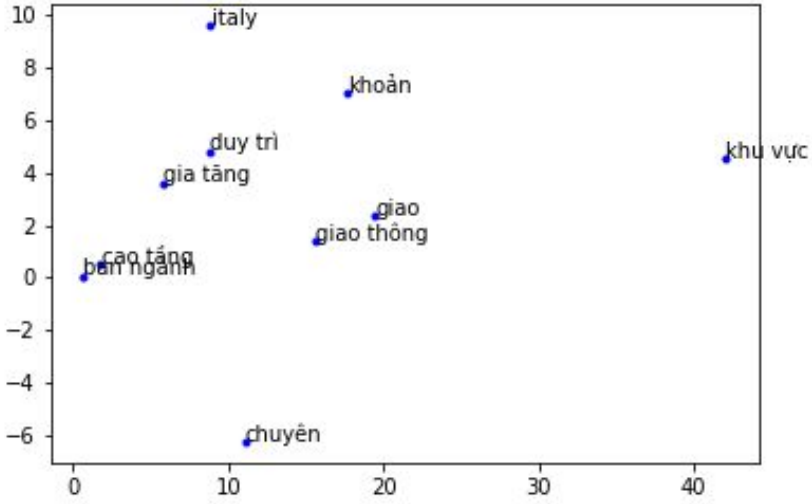

Topic 3




Topic 4


Topic 5


Topic's Wordclouds




Biểu diễn từ sử dụng lsa 2 topic (10 từ đầu)

# Nhận xét

**Điểm mạnh:**

+ Đơn giản
+ Đạt hiệu quả khá tốt
+ Khả năng ứng dụng cao

**Điểm yếu:**

- Mô hình chưa hoàn thiện
- Cần dữ liệu lớn để đạt hiệu quả

# Kết luận

- Mô hình đơn giản nhưng hoạt động khá tốt
- Dễ dùng, dễ hiểu
- Có khả năng ứng dụng nhiều
- Mô hình ngôn ngữ chưa hoàn thiện

# Tham khảo

[1] Thomas K. Landauer, Susan T. Dumais, A Solution to Plato's Problem:The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge http://lsa.colorado.edu/papers/plato/plato.annote.html

[2] Thomas K. Landauer, Danielle S. McNamara, Handbook of Latent Semantic Analysis, First Edition, Psychology Press

[3] Alex Thomo, Latent Semantic Analysis (Tutorial)
https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&cad=rja&uact=8&ved=2ahUKEwie64vhjNDIAhWMF4gKHVkaAKgQFjABegQIAhAC&url=https%3A%2F%2Fpdfs.semanticscholar.org%2F3efd%2Fa6e61747fea6b5cb5fa4f3ff0a14c86a638c.pdf&usg=AOvVaw0HGlDVpOS5yRNX-VMA1ilN

[4]Christopher D. Manning, Chapter 18: Matrix Decomposition and Latent semantic indexing, Introduction to Information Retrieval, First Edition, Cambridge University Press

....

# Cảm ơn
# thầy và các bạn đã lắng nghe!